The Malicious Technical Ecosystem: Exposing Limitations in Technical Governance of AI-Generated Non-Consensual Intimate Images of Adults

Michelle L. Ding
michelle_ding@brown.edu
Brown University
Providence, Rhode Island, USA

Harini Suresh
harini_suresh@brown.edu
Brown University
Providence, Rhode Island, USA

Abstract

In this paper, we adopt a survivor-centered approach to locate and dissect the role of sociotechnical AI governance in preventing AI-Generated Non-Consensual Intimate Images (AIG-NCII) of adults, colloquially known as "deep fake pornography." We identify a "malicious technical ecosystem" or "MTE," comprising of open-source face-swapping models and nearly 200 "nudifying" software programs that allow non-technical users to create AIG-NCII within minutes. Then, using the National Institute of Standards and Technology (NIST) AI 100-4 report as a reflection of current synthetic content governance methods, we show how the current landscape of practices fails to effectively regulate the MTE for adult AIG-NCII, as well as flawed assumptions explaining these gaps.

1 Introduction

AI-Generated Non-Consensual Intimate Images (AIG-NCII), colloquially known as "deep fake pornography," are a form of image-based sexual abuse that disproportionately harm and silence women, girls, LGBTQ+ people, and racial minorities [5, 8, 17, 29]. Image-based sexual abuse covers a spectrum of behaviors including the non-consensual *creation* of intimate images, non-consensual *distribution* of intimate images, and *threatening* to distribute those images [17, 27]. Recognizing the extensive prior work on preventing AI-generated child sexual abuse materials (AIG-CSAM) and non-consensual *distribution* of intimate images, this paper aims to contribute to the unique challenge of preventing the *creation* of AIG-NCII of *adults* [1, 4, 27, 35, 37]. These include the documented 35,545 images depicting 26 members of congress and thousands more depicting celebrities, public figures, and ordinary women and teens [3, 9, 21, 35–37]. While recent advancements in corporate image-generation models like Stability AI's Stable Diffusion have resulted in alarmingly photorealistic AIG-NCII, there has existed since 2017 a prolific ecosystem of open-source face-swapping models such as DeepFaceLab, DeepNude, and FaceSwap that support nearly 200 "nudifying" software programs allowing non-technical users to create what they call "deep fake pornography" within minutes [10, 18, 21, 26, 35, 38, 39, 42]. This decentralized technical ecosystem of open-source models and tools (which we characterize as the "malicious technical ecosystem" or "MTE") enables the creation of *AIG-NCII of adults* that are often noticeably "fake" yet *still a form of abuse* that results in mental and physical harm, reputational damage, financial costs, and disruption of social relationships [1, 8, 16, 29, 33, 40]. Image-based sexual abuse, like other forms of technology-facilitated gender-based violence, also causes a gendered chilling effect where victim-survivors retreat from online spaces [12, 17, 28]. These effects already appear in the case of AIG-NCII as seen from interviews and testimonials [8, 22].

Despite extensive documentation of the harms and scale of these models and tools, they remain largely ungoverned by current technical prevention methods. In this paper, we locate the MTE as a unique challenge for the sociotechnical governance of adult AIG-NCII. First, we identify and contextualize the MTE within the broader synthetic content pipeline of creation, distribution, and consumption. Next, by drawing on the National Institute of Standards and Technology (NIST) AI 100-4 report on governing synthetic content, we show how current governance mechanisms fail to effectively regulate the MTE: 1) *broad synthetic content governance measures* focus on transparency, which is insufficient to address the harms of watermarked, human-detectable adult AIG-NCII, 2) *AIG-NCII specific governance measures* conflate CSAM and adult NCII, proposing methods more effective in regulating CSAM than adult NCII, and 3) *adult-specific AIG-NCII governance measures* acknowledge the existence of the MTE but propose methods that will primarily regulate large corporate models. Overall, we stress the need for governance frameworks that cite AIG-NCII of adults as a high-risk harm to assess their effectiveness in regulating the MTE by considering the three listed limitations.

2 Response vs. Prevention: Locating the Role of Sociotechnical AI Governance in Mitigating AIG-NCII

Before diving into the challenges of technical prevention, it is important to contextualize the role of sociotechnical AI governance within the larger multi-stakeholder space of AIG-NCII mitigation efforts. The issue of AIG-NCII, like other forms of image-based sexual abuse and technology-facilitated gender-based violence, requires both *prevention* and *response* efforts. There are already many powerful *responses* to AIG-NCII, including: two federal legislation efforts that both passed the Senate (the DEFIANCE Act and the Take It Down Act) which require expedited platform take-down procedures and create civil liability for malicious actors; advocacy petitions for the take-down of "deepfake pornography" apps and sites; and voluntary platform trust and safety commitments [5, 20, 24, 31, 32]. While these response measures are critical to addressing the downstream distribution and consumption of AIG-NCII, they place the burden of removal on survivors, many of whom may be unwilling to report abuse, hesitant to engage with the criminal justice system, and/or lack the emotional, financial, and physical capacity to engage in lengthy take-down procedures or legal action [8, 13]. Public attitudes toward "deepfake pornography" often involve victim-blaming
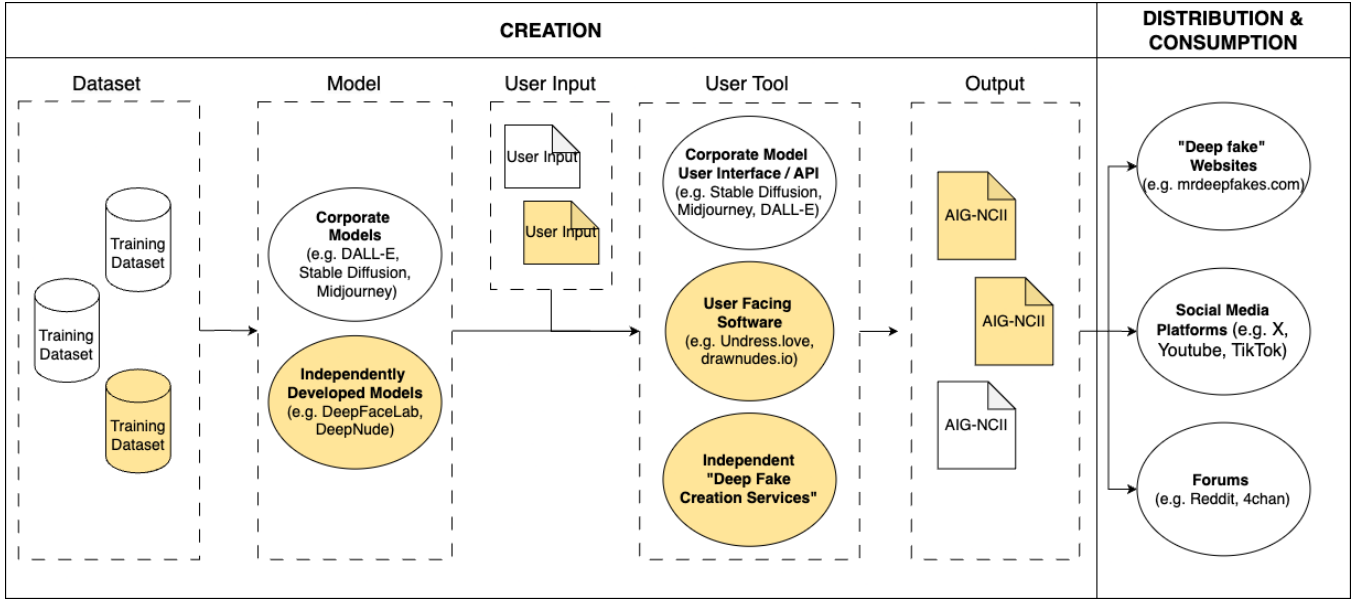
Figure 1: Supply chain for AI Generated Non-Consensual Intimate Images that expands on the NIST AI 100-4 synthetic content pipeline of creation, distribution, and consumption. Yellow circles represent the malicious technical ecosystem (MTE). Training dataset, user input, and AIG-NCII are highlighted to represent artifacts within the pipeline and not any specific example.

and stigmatization which also result in less help-seeking by victim-survivors [40]. Most importantly, responsive measures that only target the distribution fail to govern the technologies that *create* AIG-NCII, thus allowing an unimpeded stream of new AIG-NCII for survivors to take down. *Preventing the creation* of AIG-NCII, then, is a key a challenge for sociotechnical AI governance.

3 The Malicious Technical Ecosystem (MTE)

This section will characterize the MTE of adult AIG-NCII and contextualize it within a broader synthetic content pipeline. Figure 1 maps the AIG-NCII supply chain onto NIST's synthetic content pipeline of *creation, distribution,* and *consumption* [23]. As our focus is on technical prevention, we add additional granularity to *creation* by including specific artifacts such as training datasets, models, and user tools. The highlighted circles represent the core of the MTE: independently developed models, user-facing software, and independent "deep fake" creation services.

3.1 Independently Developed Models

A current quick search of "deepfake" on GitHub returns 7.2k repositories, a majority identified to generate NSFW content [25]. The most popular NCII generator is DeepFaceLab, a GAN-based face-swap model with 17.2k stars on GitHub [10, 38]. The creators of DeepFaceLab also built MrDeepFakes.com to encourage the use of their codebase [33]. This site is currently the largest dedicated "deep-fake pornography" website with over 1 billion accumulated views from 2016-2023 [21]. Other codebases that use similar architectures include FaceSwap and DeepNude [39, 42]. While some repositories like FaceSwap includes a "zero-tolerance" policy for NSFW content, there are no real technical barriers preventing AIG-NCII generation

[39]. These models have trained on various datasets including Flickr-Faces-HQ, nsfw-data-scraper, and nsfw-data-source-urls [2, 7, 34]. A study of DeepNude also found that the algorithm "cannot perform similar translations on images of men, having been specifically trained on images of women" [9].

3.2 User Facing Software and Independent "Deep Fake Creation Services"

Building upon these open-source independently developed models are over 200 dedicated AIG-NCII generation tools, or "nudifier services," that enable non-technical audiences to create AIG-NCII of any user-uploaded image within minutes [18, 21, 33]. There are also over 1700 third party paid "Deep Fake Creation Services," making up a "fully fledged online industry" [21, 29]. The outputted AIG-NCII, often watermarked, labeled "AI-generated", or captioned "not [*depicted person*]," are then distributed and consumed across various channels, including 50+ dedicated "deep fake pornography" websites, social media platforms like X, YouTube, and TikTok, and discussion forums like Reddit and 4chan [9, 21].

3.3 Supporting Technological Infrastructure

This paper is focused on technical prevention in the form of governing *models and tools*. However, the MTE is also supported by infrastructure like Github to host the codebases and chat forums (Reddit, 4chan, 8chan, and Voat) to guide new developers [6, 9, 25, 38]. Google consistently lists "deepfake creation" tools and services the at top of search results [11, 21]. Mastercard and Visa enable malicious actors to monetize their creation services [11]. Civil society organizations are also calling for these tech companies to take accountability for their support of the "MTE" [22, 25].

4 Implications of the MTE: Limitations in Current Synthetic Content Governance Practices

In this section, we anticipate three ways in which current governance practices, as reflected in the 2024 NIST report on "Reducing Risks Posed by Synthetic Content," may fail to regulate the MTE [23]. NIST begins with broadly applicable practices for digital content transparency for synthetic content, along with a specific section on technical prevention methods for AIG-NCII and AIG-CSAM that includes: 1) training data filtering, 2) input data filtering, 3) output filtering, 4) hashing confirmed AIG-CSAM and AIG-NCII, and 5) provenance data tracking techniques.

4.1 Limitation 1: Synthetic content governance measures focus on transparency, which is insufficient to address the harms of human-detectable and overtly watermarked adult AIG-NCII.

Models within the MTE like DeepFaceLab and DeepNude are not powered by state-of-the-art image-generation models trained on massive datasets. For comparison, Stable Diffusion's training dataset LAION-5B is 28,000 times larger than curated image datasets like Flickr-Faces-HQ used by DeepFaceLab [26, 30, 34, 35]. As a result, the images produced within the MTE can often be detected as "fake" by a viewer. Many tools within the MTE further watermark and label AIG-NCII outputs as "AI-generated" or "FAKE" [42]. Transparency practices like provenance data tracking, synthetic content detection, and user education labels are suitable for governing other synthetic content, such as political deep fakes and voice cloning, where harms like disinformation, election security risks or fraudulent transactions are dependent on the detectability of outputs. These efforts may also aid in accountability and content moderation during distribution and consumption of AIG-NCII. However, AIG-NCII governance measures that stop at transparency make the flawed assumption that noticeably "fake" content is no longer harmful enough to regulate. This sentiment aligns with the creator of MrDeepFakes himself, who said "I can see how some women would have psychological harm from this, but they can just say, 'It's not me, this has been faked, I can't suffer any damages from this'" [33]. On the contrary, "fake" NCII can still result in mental and physical harm, reputational damage, financial costs, and further cyber-violence [8, 22]. Most critically, "deepfake pornography," like other forms of technology-facilitated gender based violence, creates a gendered chilling effect that disproportionately silences gender, racial, and sexual minorities [14]. A study on public attitudes further found that "labeling pornographic deepfakes as fictional did not mitigate the videos' perceived wrongfulness" [15].

4.2 Limitation 2: AIG-NCII specific measures conflate CSAM and adult NCII, proposing methods more effective in regulating CSAM than adult NCII.

Adult AIG-NCII and CSAM are often conflated into one category by NIST and other multistakeholder frameworks [5, 24]. However, examples of effective prevention methods primarily target CSAM. For example, the key study cited in "Method 1: Training Data Filtering" uses Microsoft's PhotoDNA tool to query existing databases of known illegal CSAM images [19, 35]. With CSAM, consent is not debatable: any sexually explicit depiction of a minor is illegal under federal child pornography law [41]. Thus, AIG-CSAM tools can draw upon existing law enforcement databases of CSAM. With adults, it is more difficult to collect such a database as consent is not automatically reflected in the dataset. Although CSAM and adult-NCII methods may be similar, it is important for any governance framework or method that aims to be applicable to *both* CSAM and adult-NCII to consider this distinction.

4.3 Limitation 3: Adult-specific AIG-NCII governance measures acknowledge the existence of the MTE but propose methods that will only regulate large corporate models.

While NIST also references the same models and tools described in our MTE, their governance methods for adult NCII primarily apply to the architecture of larger corporate image-generation models. For example, "Input Data Filtering" methods like keyword blocking and "Red-Teaming and Testing" methods like adversarial prompt generation assume a prompt-based image-generation model. Therefore, these methods are inapplicable for the MTE's "deep fake pornography creation tools" where user input is just the photo they want to "nudify." The gap in governance here is primarily caused by the assumption that the image-generation technology is trustworthy and NCII is a result of malicious users. In the MTE, the *technology itself* is malicious, developed intentionally for a single goal of generating NCII. Frameworks or methods that aim to prevent AIG-NCII should consider their applicability to the "MTE" *in addition* to corporate models.

5 Looking Forward: Towards Survivor-Centered Prevention of Adult AIG-NCII

Survivors and advocates have and continue to fight long and difficult battles against image-based sexual abuse and technology-facilitated gender-based violence. This paper builds upon their tremendous work by calling upon the sociotechnical research community to prioritize the improvement of current governance methods that fall short in effectively regulating the malicious technical ecosystem that produces tens of thousands of life-threatening adult AIG-NCII reaching billions of viewers every year. Under the MTE paradigm, a harmful image can be reported and removed or a particular "deepfake" repository can be taken down, but just as quickly, new ones can spring up. In this paper, we point out three ways in which current synthetic content governance methods fail to govern the MTE producing adult AIG-NCII content. More importantly, we expose some *flawed fundamental assumptions* behind these governance methods. The MTE and its terrible consequences show us that fake content is *still* harmful content worthy of technical governance, and that harmful outputs are not just products of malicious users, but of malicious technologies themselves.

References

[1] Shelby Akerley. 2021. Let's Talk About (Fake) Sex Baby: A Deep Dive Into the Distributive Harms of Deepfake Pornography. *Arizona Law Journal of Emerging Technologies* 4, 1 (March 2021). doi:10.2458/azlawjet.5505

[2] Alex Kim. 2019. nsfw_data_scraper. https://github.com/alex000kim/nsfw_data_scraper

[3] American Sunlight Project. 2024. Deepfake Pornography Targeting Members of Congress. https://www.americansunlight.org/updates/deepfake-pornography-targeting-members-of-congress

[4] Becca Branum. 2024. NDII Victims Deserve Help. Let's Build an Effective Take-down System. https://cdt.org/insights/ndii-victims-deserve-help-lets-build-an-effective-takedown-system/

[5] Center for Democracy and Technology. 2025. IBSA Principles. https://ibsaprinciples.org/ Signatories: CDT, CCRI and NNEDV include: Aylo, Bumble, Google, Open AI, Match, Meta, Microsoft, Reddit, Free Speech Coalition, the Human Rights Campaign, LGBT Tech, the Reclaim Coalition, the Sexual Violence Prevention Association, the UC Center for Race and Digital Justice, Professor Elissa Redmiles, and Dr. Lucy Qin..

[6] Cinecom.net. 2019. DEEPFAKE Tutorial: A Beginners Guide (using DeepFace Lab). https://www.youtube.com/watch?v=t59gRbpYMiY

[7] Evgeny Bazarov and Jean-Philippe Paradis. 2019. nsfw_data_source_urls. https://github.com/EBazarov/nsfw_data_source_urls

[8] Asher Flynn, Anastasia Powell, Adrian J Scott, and Elena Cama. 2022. Deep-fakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging Form of Image-Based Sexual Abuse. *The British Journal of Criminology* 62, 6 (Oct. 2022), 1341–1358. doi:10.1093/bjc/azab111

[9] Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. 2019. *The State of Deepfakes: Landscape, Threats, and Impact.* Technical Report. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

[10] iperov. 2018. DeepFaceLab. https://github.com/iperov/DeepFaceLab

[11] Kat Tenbarge. 2023. Found through Google, bought with Visa and Mastercard: Inside the deepfake porn economy. *NBC News* (March 2023). https://www.nbcnews.com/tech/internet/deepfake-porn-ai-mr-deep-fake-economy-google-visa-mastercard-download-rcna75071

[12] Kristine Baekgaard. 2024. *Technology-Facilitated Gender-Based Violence: An Emerging Issue in Women, Peace and Security.* Technical Report. Georgetown In-stitute for Women, Peace and Security. https://giwps.georgetown.edu/resource/technology-facilitated-gender-based-violence/

[13] Katherine Lorenz, Anne Kirkner, and Sarah E. Ullman. 2019. A Qualitative Study Of Sexual Assault Survivors' Post-Assault Legal System Experiences. *Journal of Trauma & Dissociation* 20, 3 (May 2019), 263–287. doi:10.1080/15299732.2019.1592643

[14] Sophie Maddocks. 2020. 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies* 7, 4 (Oct. 2020), 415–423. doi:10.1080/23268743.2020.1757499

[15] Matthew B. Kugler and Carly Pace. 2021. Deepfake Privacy: Attitudes and Regulation. *Northwestern University Law Review* 116, 3 (2021). https://scholarlycommons.law.northwestern.edu/nulr/vol116/iss3/1

[16] Clare McGlynn, Kelly Johnson, Erika Rackley, Nicola Henry, Nicola Gavey, Asher Flynn, and Anastasia Powell. 2021. 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse. *Social & Legal Studies* 30, 4 (Aug. 2021), 541–562. doi:10.1177/0964663920947791

[17] Clare McGlynn, Erika Rackley, and Ruth Houghton. 2017. Beyond 'Revenge Porn': The Continuum of Image-Based Sexual Abuse. *Feminist Legal Studies* 25, 1 (April 2017), 25–46. doi:10.1007/s10691-017-9343-2

[18] Pulak Mehta, Gauri Jagatap, Kevin Gallagher, Brian Timmerman, Progga Deb, Sid-dharth Garg, Rachel Greenstadt, and Brendan Dolan-Gavitt. 2023. Can Deepfakes be created on a whim?. In *Companion Proceedings of the ACM Web Conference 2023.* ACM, Austin TX USA, 1324–1334. doi:10.1145/3543873.3587581

[19] Microsoft. 2015. PhotoDNA. https://www.microsoft.com/en-us/photodna

[20] My Image My Choice. [n. d.]. #BlockMrDeepfakes and the 9500+ websites dedicated to image-based sexual abuse. https://www.change.org/p/shut-down-mrdeepfakes-and-websites-dedicated-to-image-based-sexual-abuse

[21] My Image My Choice. 2023. The Dataset! https://docs.google.com/spreadsheets/d/1vM-J_9NvOc4xHKbqqlZWzpQ0a0zsrDX-7Jx-wPn1hwg/edit?usp=sharing

[22] My Image My Choice. 2024. DEEPFAKE ABUSE: LANDSCAPE ANALYSIS: The Exponential Rise of Deepfake Abuse in 2023 - 2024. https://www.canva.com/design/DAGLHpt6WlY/Hfztqtw_-tKza_2l1cPNrA/view?utm_content=DAGLHpt6WlY&utm_campaign=designshare&utm_medium=link&utm_source=editor#1

[23] Gaithersburg Md Nist. 2024. *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency.* Technical Report NIST AI NIST AI 100-4. National Institute of Standards and Technology, Gaithersburg, MD. NIST AI NIST AI 100–4 pages. doi:10.6028/NIST.AI.100-4

[24] Partnership on AI. 2023. *Responsible Practices for Synthetic Media A Framework for Collective Action.* Technical Report. https://syntheticmedia.partnershiponai.org/

[25] Patrick Trueman, Esq. and Dawn Hawkins. 2023. GitHub hosting source code for sexually exploitative technology, facilitating image-based sexual abuse (IBSA), sexual exploitation, and promoting the dangerous use of generative-AI. https://endsexualexploitation.org/wp-content/uploads/Microsoft-GitHub-Notification-Letter_DDL-2023_FINAL.pdf

[26] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2020. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. doi:10.48550/ARXIV.2005.05535 Version Number: 5.

[27] Lucy Qin, Vaughn Hamilton, Sharon Wang, Yigit Aydinalp, Marin Scarlett, and Elissa M. Redmiles. 2024. "Did They F***ing Consent to That?": Safer Digital Intimacy via Proactive Protection Against Image-Based Sexual Abuse. doi:10.48550/arXiv.2403.04659 arXiv:2403.04659 [cs].

[28] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. 2019. "They Don't Leave Us Alone Anywhere We Go": Gender and Digital Abuse in South Asia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM, Glasgow Scotland Uk, 1–14. doi:10.1145/3290605.3300232

[29] Santiago Lakatos. 2023. *A Revealing Picture.* Technical Report. Graphika. https://graphika.com/reports/a-revealing-picture

[30] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. doi:10.48550/ARXIV.2210.08402 Version Number: 1.

[31] Sen. Cruz, Ted [R-TX]. 2023. TAKE IT DOWN Act. https://www.congress.gov/bill/118th-congress/senate-bill/4569/cosponsors

[32] Sen. Durbin, Richard J. [D-IL]. 2023. DEFIANCE Act of 2024. https://www.congress.gov/bill/118th-congress/senate-bill/3696/text

[33] Sophie Cockerham. 2022. Deepfake porn wrecks lives – but, as one woman discovered, it takes just 8 seconds to make an image. *METRO* (Oct. 2022). https://metro.co.uk/2022/10/21/deepfake-porn-takes-seconds-to-make-but-can-wreck-lives-17605029/

[34] Tero Karras and Janne Hellsten. 2019. Flickr-Faces-HQ Dataset (FFHQ). https://github.com/NVlabs/ffhq-dataset

[35] David Thiel. 2024. Identifying and Eliminating CSAM in Generative ML Training Data and Models. (2024). doi:10.25740/KH752SM9123 Publisher: Stanford Digital Repository.

[36] Thorn & WeProtect Global Alliance. 2024. *Evolving Technologies Hori-zon Scan: A review of technologies carrying notable risk and opportu-nity in the fight against technology-facilitated child sexual exploitation.* Technical Report. https://info.thorn.org/hubfs/Research/Thorn_x_WPGA_EvolvingTechnologies_Dec2024.pdf

[37] Thorn, All Tech Is Human. 2024. *Safety by Design for Generative AI: Preventing Child Sexual Abuse.* Technical Report. https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf Signatories: AWS AI, Civitai, Hugging Face, Inflection, Metaphysic, Stability AI, Teleperformance.

[38] Brian Timmerman, Pulak Mehta, Progga Deb, Kevin Gallagher, Brendan Dolan-Gavitt, Siddharth Garg, and Rachel Greenstadt. 2023. Studying the Online Deepfake Community. *Journal of Online Trust and Safety* 2, 1 (Sept. 2023). doi:10.54501/jots.v2i1.126

[39] torzdf. 2017. FaceSwap. https://github.com/deepfakes/faceswap

[40] Rebecca Umbach, Nicola Henry, and Gemma Beard. 2025. Prevalence and Impacts of Image-Based Sexual Abuse Victimization: A Multinational Study. doi:10.48550/arXiv.2503.04988 arXiv:2503.04988 [cs].

[41] U.S. Department of Justice. 2023. Citizen's Guide To U.S. Federal Law On Child Pornography. https://www.justice.gov/criminal/criminal-ceos/citizens-guide-us-federal-law-child-pornography

[42] yuanxiao. 2020. DeepNude. https://gitlab.com/ai-image-and-text-processing/DeepNude-an-Image-to-Image-technology/-/tree/master/DeepNude_software_itself